REMARKS

I. The Prior Art Rejections

Claims 8, 11-15, 17-23, and 26-29, all the claims pending in the application, stand rejected under 35 U.S.C. §103(a) as being unpatentable over Lantrip et al., hereinafter "Lantrip"(U.S. Patent No. 6,298,174) in view of Ruocco et al., hereinafter "Ruocco" (U.S. Patent No. 5,864,855). Applicants respectfully traverse these rejections because neither Ruocco, nor Lantrip, teach or suggest **a method or system which involves generating centroid seeds based on document classes from one dataset and using those centroid seeds when clustering documents in a new and different, but related, dataset**.

A. The Cited Prior Art.

1. Lantrip discloses a method of visually displaying the relative content of a large number of documents (see Abstract). Specifically, the relationships of the documents are presented in a 3D landscape with the relative sizes and heights of peaks in the landscape representing the relative significance of a relationship of an individual document in the document set to a topic or term (see col. 2, lines 26-31). As discussed in col. 2, lines 30-55, the Lantrip method includes the steps of creating a vector for each document in the set such that each vector represents the relative relationship of an individual document to a term or topic. Then, the vectors are arranged into clusters. For each cluster, the "centroid coordinates" (i.e., the coordinates for the center of the cluster mass) are determined as well as the distance of each document in a cluster from the centroid. This information is ultimately used to generate the 3D display. Thus, Lantrip is only concerned with a single set of documents.

2. Ruocco discloses a processing system that utilizes parallel processors for organizing and clustering a large number of documents (see Abstract). Col. 4, lines 35-45, of Ruocco summarizes conventional document clustering in which text documents are converted to document vectors and clustered. Clusters of documents are represented by cluster vectors. As

each new document is processed, its document vector is compared to the cluster vector for each cluster. If a document vector is similar to a given cluster vector, it is placed in that cluster. If it is not, a new cluster is formed. Col. 4, lines 45-60, summarizes the Ruocco method that uses multiple processors to store various clusters. In the Ruocco method, as each new document in a dataset is processed, its document vector is simultaneously compared to P cluster vectors, where P is the number of processors within the system. That is, the processors work simultaneously and each processor compares the document vector to each of its clusters.

B. Rejection Of Independent Claim 8, 15, 20 and 23 Based

The Applicants submit that neither Ruocco, nor Lantrip, teach or suggest the following features of independent claim 8 or the similar features found in independent claims 15, 20 and 23: (1) "wherein said cluster generator clusters second documents in said second dataset using said centroid seeds such that said second dataset has a similar, based on said centroid seeds, clustering to that of said first dataset"; and (2) "wherein said second dataset comprises a new, but related, based on said centroid seeds, dataset different than said first dataset."

More specifically, in rejecting claim 8, 15, 20 and 23, the Office Action provides that Lantrip discloses "creating centroid seeds based on said first document classes (in col. 2, lines 43-45, the invention finds centroids)." The Applicants respectfully disagree. Lantrip does not disclose creating centroid seeds, as indicated in the Office Action, but rather col. 2, lines 43-45 of Lantrip refers to calculating "centroid corrdinates" of clusters in a dataset (i.e., determining the coordinates for the center mass of each cluster after all the clusters are formed). Thus, Lantrip discloses centroid coordinates, but not "centroid seeds." The centroid coordinates of Lantrip are mapped points in a cluster that are used by Lantrip to ultimately generate a 3D display representing the dataset. They are not "seeds" that are subsequently used as the starting points to create clusters for a new and different, but related, dataset.

As explained in lines 5-20 of the specification of the present invention, each "centroid seed" is an average value of all the values in a document class (i.e., as defined by a cluster) from one dataset and each of these seeds are subsequently used to generate an initial cluster in another dataset. That is, conventionally, the starting points for creating clusters in a dataset are selected in some random manner. However, the present invention intentionally biases the clustering algorithm towards the classes in a first dataset by using centroid seeds generated from those first classes as the starting points for the clusters of a second new and different, but related, dataset.

The Office Action further acknowledges that Lantrip does not disclose "clustering second documents in a second dataset using said centroid seeds." Therefore, the Office Action provides that "in col. 14, lines 10-45 of Ruocco, Ruocco discloses in the claim processing in parallel second datasets based on cluster information from previous cluster vectors (see col. 14, lines 28-30) in order to gain the benefit of information from previous clusters to improve analysis of datasets. Ruocco's invention further may be interpreted such that said second dataset has a similar clustering to that of said first dataset (as the term "similar" is sufficiently broad that any two given datasets would have some degree of similarity, see 35 U.S.C. 112 rejection, above.), further wherein said second data set comprises a new, but related dataset different than said first dataset (once the first dataset is transformed, it is by definition, a new, but related dataset). It would have been obvious to one of ordinary skill in the art at the time of the invention to use the information contained in the centroid seeds from Lantrip for subsequent datasets as in Ruocco in order to improve analysis of subsequent datasets." The Applicants respectfully disagree.

Ruocco only discloses clustering documents in a single dataset. It does not teach or suggest clustering first documents in a first dataset and then subsequently clustering second documents in a second dataset using centroid seeds generated based on the first document clusters (i.e., based on the first document classes). Specifically, the cited portion of Ruocco discloses a computer system having parallel processors and a method which individually examines documents and assigns similar documents to a particular cluster that is in turn assigned to a particular processor. The method is initiated by selecting a first document, converting it into a first document vector, designating the first document vector as a first cluster vector and assigning the first cluster vector to a first processor (see col. 14, lines 18-27). Then, a second document is selected and converted into a second document vector. This second document vector is compared to the first document vector. If they are similar the second document vector is assigned to the first cluster in the first processor. If they are not, the second document vector is assigned to a second processor (see col. 14, lines 27-35).

Thus, Ruocco simply discloses an improvement over legacy methods of clustering documents in a single dataset (e.g., a 32 document set, a 64 document set, a 128 document set (see col. 7, lines 30-35). That is, due to a determined need to organize a large number of documents in a document set, the Ruocco method forms clusters of similar documents, but instead of maintaining those clusters on a single processor, Ruocco maintains one or more clusters on multiple different processors. The initial cluster vector for a given cluster in Ruocco is disclosed as being the first document vector in the cluster. After that it is a mathematical average of document vectors in the cluster (see col. 14, line 25-26). As each new document is added to the dataset, the document is assigned to a particular cluster on a particular processor based on similarities between its document vector and the cluster vector of the particular cluster (see col. 14, lines 28-46).

No where in the cited portion of Ruocco does it teach or disclose "processing in parallel second datasets based on cluster information from previous cluster vectors in order to gain the benefit of information from previous clusters to improve analysis of datasets," as indicated by the Office Action. While Ruocco uses cluster vectors (i.e., either the document vector if there is only one document in a cluster or the mathematical average of document vectors in a cluster of more than one document) when determining whether or not to add a new document to an old cluster or to create a new cluster in a dataset, Ruocco does not teach or suggest using any information from the first document classes (i.e., first clusters) of a first dataset when clustering second documents in a second dataset. That is, presumably, for each new dataset (i.e., for each new set of documents) in Ruocco, the method would follow the exact same steps as disclosed in the claim set in column 14 such that the first document vector will be designated as the first cluster vector and so on.

Furthermore, nothing in Lantrip or Ruocco teaches or suggests the desirability of the present invention. The Office Action indicates that it "would have been obvious to one of ordinary skill in the art at the time of the invention to use the information contained in the centroid seeds from Lantrip for subsequent datasets as in Ruocco in order to improve analysis of subsequent datasets." However, as discussed above, Lantrip does not disclose "centroid seeds", but rather calculating centroid coordinates for clusters of documents in a dataset in order to generate a 3D display of the dataset. Thus, a combination of Lantrip and Ruocco would result in the calculation of all centroid coordinates for all of the clusters stored on all of the processors in order to generate a 3D display.

Additionally, also as discussed above, the starting points for creating clusters in a dataset are conventionally selected in some random manner. In Ruocco, the starting point is based on the first document vector and the document vectors of any subsequently processed document that is not similar to already established clusters. In Landript, the starting point is not disclosed. Contrarily, the Applicants recognized that clustering documents from new and different, but related datasets, results in a lack of continuity that is a drawback when one in interested in tracking changes in the data (trends) over time (see page 5, lines 0-9). This is a problem not addressed in either Ruocco or Landrip (as they disclose conventional techniques for selecting the starting points and no technique, respectively). The present invention solves this problem by intentionally biasing the clustering algorithm towards first classes in a first dataset by using centroid seeds generated from clusters corresponding to those first classes as the starting points for clustering a second new and different, but related, dataset. More specifically, the claimed invention solves the problem of finding new categories in a second data set that did not exist in the first data set, while at the same time maintaining as nearly as possible categories from the first data set as categories in the second data set. With the claimed invention, there is no requirement, and in fact it is not assumed, that the first and second datasets have any of the same data elements in them. They are allowed to have some of the same elements, but this is in no way a requirement for the claimed invention. The claimed invention is designed in such a way as to find the similarities between the two datasets, where they exist, while at the same time finding the key differences (emerging concepts) in the second dataset.

Therefore, the Applicants submit that the applied prior art references alone and in combination do teach or suggest the following features of independent claim 8 or the similar features found in independent claims 15, 20 and 23: (1) "wherein said cluster generator clusters second documents in said second dataset using said centroid seeds such that said second dataset has a similar, based on said centroid seeds, clustering to that of said first dataset"; and (2) "wherein said second dataset comprises a new, but related, based on said centroid seeds, dataset different than said first dataset." Consequently, independent claims 8, 15, 20, and 23 are patentable over the applied prior art references. Further, dependent claims 11-14, 17-19, 21, 22, and 26-29 are similarly patentable, not only by virtue of their dependency from a patentable claim, but also by virtue of the additional features of the invention they define. In view of the foregoing, the Examiner is respectfully requested to reconsider and withdraw this rejection.

II. Formal Matters and Conclusion

In view of the foregoing, Applicants submit that claims 8, 11-15, 17-23, and 26-29, all the claims presently pending in the application, are patentably distinct from the prior art of record and are in condition for allowance. The Examiner is respectfully requested to pass the above application to issue at the earliest possible time.

Should the Examiner find the application to be other than in condition for allowance, the Examiner is requested to contact the undersigned at the local telephone number listed below to discuss any other changes deemed necessary.

Please charge any deficiencies and credit any overpayments to Attorney's Deposit Account Number 09-0441.

Respectfully submitted,

Dated: May 25, 2007

/Pamela M. Riley/
Pamela M. Riley
Registration No. 40,146

Gibb & Rahman, LLC
2568-A Riva Road, Suite 304
Annapolis, MD 21401
(410) 573-0227
Customer Number: 29154